Underlying elements of image quality assessment: : Preference and terminology for communicating image quality characteristics

Virtanen, Toni

2020-03-04

Virtanen , T , Nuutinen , M & Häkkinen , J 2020 , ' Underlying elements of image quality assessment: Preference and terminology for communicating image quality characteristics ' , Psychology of Aesthetics, Creativity and the Arts . https://doi.org/10.1037/aca0000312

http://hdl.handle.net/10138/335046
https://doi.org/10.1037/aca0000312

unspecified
acceptedVersion

Underlying elements of image quality assessment: Preference and terminology for communicating image quality characteristics

Toni Virtanen, Mikko Nuutinen, and Jukka Häkkinen

University of Helsinki

Author Note

Toni Virtanen, Mikko Nuutinen, and Jukka Häkkinen, Department of Psychology, Faculty of Medicine, University of Helsinki.

This study was supported in part by Doctoral School of User-Centered Information Technology (UCIT). The authors also want to thank Nokia Co. and Microsoft Co. for providing the test materials for the studies.

Correspondence concerning this article should be addressed to Toni Virtanen, Department of Psychology, Faculty of Medicine, University of Helsinki. PO Box 21 (Haartmaninkatu 3), 00014 University of Helsinki, Finland. E-mail: toni.virtanen@helsinki.fi

Abstract

Image quality markedly affects the evaluation of images, and its control is crucial in studies using natural visual scenes as stimuli. Various image elements, such as sharpness or naturalness, can impact how observers view images and more directly how they evaluate their quality. To gain a better understanding of the types of interactions between these various elements, we conducted a study with a large set of images with multiple overlapping distortions, covering a wide range of quality variation. Observers assigned a quality rating on a 0-10 scale plus a verbal description of the images, explaining the elements on which their rating was based. Regression model predicting image quality ratings using 68 attributes uncovered the link between verbal descriptions and quality ratings and the importance of the image quality rating for each of the 68 image attributes. Brightness, naturalness, and good colors seem to be related to the highest image quality preference. However, the most important elements for predicting good image quality were related to image fidelity such as graininess and sharpness. This indicates that a certain level of image fidelity must be achieved before more subjective associations with, for instance, naturalness can emerge. Of the attributes, 72% had a negative impact on the preference judgment. This negative bias may be due to the fact that there are more ways that observers can perceive an image to fail than to excel when they are asked to evaluate image quality.

Keywords: Aesthetic preference, Image quality, Attribute, Mean opinion score, mixed method

Images are ubiquitous; whether they are on billboards, art galleries, web sites, device screens, portraits, or family albums, they are simply part of our lives. As cameras have become a standard feature in mobile phones, almost everyone also has the experience of taking photographs themselves.

Image quality can be seen as a subsection of the wider multidisciplinary field of Quality of Experience (QoE), and even the research within image quality is multidisciplinary, including disciplines such as vision science (To, Lovell, Troscianko, & Tolhurst, 2008), color science (S. N. Yendrikhovskij, De Ridder, Fedorovskaya, & Blommaert, 1997), computational sciences (Dodge & Karam, 2019; J. A. Redi, Zhu, Ridder, de, De Ridder, & Heynderickx, 2015), and behavioral sciences (Augustin, Wagemans, & Carbon, 2012; L. Leder, Belke, Oeberst, & Augustin, 2004; T. Leisti, Radun, Virtanen, Halonen, & Nyman, 2009; Tinio, Leder, & Strasser, 2011).

Probably the most ambitious effort to create a comprehensive definition of Quality of Experience (QoE) has been given by Qualinet, the European network on Quality of Experience in multimedia systems and services, where a working definition of QoE was created by 49 researchers representing 18 different European countries: *"Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state."*(Callet, Möller, & Perkins, 2012). The researchers themselves point out that the current definition does not address the degree of success achieved by the artist in conveying the intended message, but rather the influence of the technical system or processing on the artist's work. Image quality has been shown to have a great influence on the evaluation of images, and its control is crucial in investigations using natural visual scenes as stimuli (Tinio & Leder, 2009). But why are certain images preferred over others?

One definition of image quality is related to image fidelity, particularly to perceptual fluency (Reber, Schwarz, & Winkielman, 2004). Images with clear perceptual fluency are preferred more as they can convey the message better and are easier to interpret by the viewer. The aesthetic pleasure of viewing any abstract art at first hand seems to contradict this processing fluency theory. Graf & Landwehr (2015), however, introduced a dual-processing perspective to processing fluency, where abstract art, with its low processing fluency, would introduce aesthetic pleasure through cognitive enrichment, while natural scene images would be processed mostly at an automatic level, where clear processing fluency would be preferred. It has also been speculated that people have an understanding of images being merely a representation of the scene and show a preference for ones that possess high image quality and artistic value (Tinio et al., 2011). Moreover, being exposed to countless images during their lifetime, people become accustomed to evaluating image quality and have certain expectations of what constitutes good image quality. However, as imaging devices develop and new technologies emerge, expectations will change as well.

Another approach to defining image quality is conceptualizing it as a combination of various elements such as sharpness and colorfulness. These elements are weighted and summed to create the overall model of image quality. Several models for how this summation is created have been suggested (Bech et al., 1996; Engeldrum, 2004; IEEE, 2007; Janssen & Blommaert, 1997; Karel Fliegel, Christian Timmerer, 2013; Keelan, 2002). This definition has the benefit of combining views from multidisciplinary stakeholders approaching image quality from different directions. The summation and weighting of the elements can be viewed as a cognitive-affective process of the viewer. Berlyne (1972) suggested that preference is formed from the combination of pleasantness, interest, liking, and complexity. Four algorithmic measures: *visual simplicity, visual symmetry, visual contrast, and visual self-similarity,* which extract low-level visual stimulus properties predictive of aesthetic

preferences according to the processing fluency theory have also been suggested (Mayer & Landwehr, 2018). A multidimensional scaling (MDS) study by O'Hare and Gordon (1977) linked Realistic-Unrealistic, Clear-Indefinite, and Symmetrical-Asymmetrical dimensions to the preference of paintings.

The concept of the summation of image elements and scene statistics is also used from the technological point of view when developing image quality assessment algorithms or other methods to quantify QoE. For example, ACQUINE (Aesthetic Quality Inference Engine) is a machine learning-based system that showcases computer-based prediction of aesthetic quality for natural color photographic images (Datta & Wang, 2010). The Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE), in turn, is a natural scene statistic-based distortion-generic blind/no-reference (NR) quality assessment algorithm (Mittal, Moorthy, & Bovik, 2012). The Video (BLIINDS) uses a similar natural scene statistic model-based approach to the no-reference/blind video quality assessment problem (Saad, Bovik, & Charrier, 2010). OSCAR (On-site Composition and Aesthetics Feedback), on the other hand, focuses on compositional aspects (Yao, Suryanarayan, Qiao, Wang, & Li, 2012). Although the examples are widely different, all approaches utilize some sort of summation of image elements to create a model for image quality, QoE, or even aesthetic preference.

Perhaps because of its multidisciplinary relevance, the terminology of image quality has been poorly defined (Augustin et al., 2012; Virtanen, Nuutinen, & Häkkinen, 2019). Reference wheels and terminology lexicons have a long tradition in the sensory evaluation fields (Lawless & Civille, 2013), such as taste (Chen, Rhodes, Crawford, & Hambuchen, 2015; Gawel, Oberholster, & Francis, 2000; Lawless, Hottenstein, & Ellingsworth, 2012; Meilgaard, Dalgliesh, & Clapperton, 1979) and smell (Zarzo & Stanton, 2009), where they are used to standardize terminology and facilitate communication between various stakeholders. Similar approaches for color prints (Pedersen, Bonnier, Hardenberg, &

Albregtsen, 2010) and, more recently, for image quality in general (Virtanen et al., 2019) have been presented. An even more general foundation for terminology of aesthetic word use has been given by Augustin et al. (2012), who examined the word use with eight different object classes, showing an interplay between generality and specificity in aesthetic word usage. Another method to assess any type of visual art with perceptual and conceptual attributes was provided in the form of the Assessment of Art Attributes (AAA), where the six attributes of *Balance, Color Saturation, Color Temperature, Depth, Complexity, and Stroke Style* refer to perceptual properties and the six attributes of *Abstractness, Animacy, Emotionality, Realism, Representational Accuracy, and Symbolism* refer to conceptual properties (Chatterjee, Widick, Sternschein, Smith, & Bromberger, 2010).

A study by Nyman et al. (2010) demonstrated how images with a low quality rating were characterized by different terminology than images with a high quality rating, suggesting a paradigm shift in the subjective decision-making space as a function of preference. In other words, images with low quality are evaluated with a different set of rules and terms than images with high quality. Another study with printed images further classified the terminology on image quality as having two levels: low level and high level (see Fechner, 1876; H. Leder, 2014). The most important low-level attributes were brightness of color, sharpness, graininess, brightness, color quality, gloss, contrast, and lightness. The high-level attributes, in turn, were used to funnel the importance of the low-level attributes and consisted of realism, naturalness, clarity, depth, and aesthetic associations (T. Leisti et al., 2009).

This study explored how the word use changes for high-quality images and low-quality images and examined the interplay between image quality ratings and word use with multiply distorted images. This study presents a large set of images covering a wide range of quality variation. The images were rated on a 0-10 scale and given free verbal descriptions of the

most decisive features related to the image. Altogether 68 attributes were derived from the free descriptions using a method described in an earlier study where 39,415 quotations from 146 observers evaluating 62 scenes were categorized according to a descriptive terminology lexicon, the Image Quality wheel (Virtanen et al., 2019). This study builds on the earlier work by creating a regression model to predict image quality ratings using the 68 attributes, uncovering a link between verbal descriptions and quality ratings and the importance of image quality rating for each of the 68 attributes.

Methods

Observers

A total of 59 (95% female) native Finnish-speaking participants were recruited through the University of Helsinki student mailing lists using a simple opportunity sampling method. Neither age nor gender was controlled, although differences in linguistic behavior in social contexts may exist between genders and age groups. However, the solitary task of assessing and writing free descriptions about the images should not introduce these kinds of social roles and differences. Participants received a movie ticket as remuneration for their time. None of the participants were professionally involved in photography. All participants had their near visual acuity (ETDRS chart, Precision Vision Inc.), near contrast vision (Near F.A.C.T., Stereo Optical Inc.), and color vision (Farnsworth D-15) tested prior to the experiments. Normal or corrected-to-normal vision was a requirement for participation.

Experiments were separated into four studies (Table 1). Study 1 had six different contents that were manipulated by 13 different Image Signal Processes; thus, the total number of evaluated images was 6 x 13 = 78. After Study 1, the observers provided feedback that rating 13 images from a single scene six times in a row was tedious. Therefore, for the later studies we increased the number of different contents to eight and decreased the number of

manipulations (ISP) to nine, keeping the number of images similar to that of Study 1, but providing more variability in scenes. We had 15 participants in each study apart from study 4, which had 14 participants. Participants were mostly young females with an average age of 23 years. The duration of the experiment overall was 1.5 hours, including vision tests, instructions, practice, and possible breaks.

Table 1

Experiments were divided into four studies. In each study, the different contents depicting different scenes were manipulated by different Image Signal Processing steps. For example, 8 x 9 creates 72 different images to rate. Observers were young adults with a mean age of 23 years. The average duration of the experiment overall was 1.5 hours, including vision tests, instructions, practice, and possible breaks

Study	Contents	Image signal processing	Images	Observers	Average Age	Average Duration
1	6	13	78	15	22	1:32
2	8	9	72	15	23	1:34
3	8	9	72	15	23	1:29
4	8	9	72	14	24	1:37
Total	30	40	294	59	23	1:33

Materials

Images were shot using three imaging devices of the same model that were passed around to different individuals to gather as many different images as possible. The raw signal was then processed using 40 different image signal processing (ISP) pipes. The ISP affects exposure, noise reduction, sharpening, and white balance algorithms, among others. To give

an example of what the ISP does, we can consider how color images are created. The imaging sensor is just a color agnostic array of photosensitive pixels. To get a color image, the pixels in the imaging sensor are behind a Bayer color filter of Red/Green/Blue/Green array. ISP does the demosaicing of this array and guesses the right color hue for each pixel based on the information on the varying intensity of nearby pixels under the mosaic color filter. In other words, ISPs are basically the in-device darkroom that processes the raw signal from the sensor, creating the visual image that we actually see on our device screen. Every digital camera does this before we even notice that any manipulation has been done to the image. In our image samples, it would be the same as if the images had been taken with different cameras that had identical lens and sensor characteristics. An ISP is always a compromise between computing power, battery consumption, speed, and quality of the original source signal from the sensor. The images were processed by ISPs created by different vendors, each competing to generate a combination of processing that would produce images with the most pleasing end result. The variation in quality comes from the choices and differences in signal processing thresholds. For example, how much noise is allowed before the de-noising algorithm starts to reduce it? Then again, how much blur is the de-noising allowed to add to the noisy image so that it is still considered an enhancement? And finally, could some sharpening algorithm be used to offset the effect of the blur from de-noising and how much sharpening is too much? These are basically all aesthetic choices pre-made by the manufacturer of which the end user is rarely aware. Many Professional DSLR cameras allow this processing to be turned off or modified, but demosaicing is basically a mandatory step in the signal processing and potential tweaks are coarse compared with the precision in the manufacturer-set thresholds. Unfortunately, the ISPs are property of our industry partners and we cannot disclose the details about the selections and thresholds used to develop the actual visual image from the raw signal output from the imaging sensor. On the other hand, had we

used any of the multiple cameras available on the market, the same limited amount of information would be provided on the signal processing for those devices. The images in this study represent actual photographs that would be produced by different cameras having identical lens and sensor characteristics.

Previous studies have shown that image distortions can influence the way that people view images (Radun, Nuutinen, Leisti, & Häkkinen, 2016; J. Redi, Liu, Zunino, & Heynderickx, 2011). The ISP manipulations fit well for this type of study, as multiple overlapping manipulations, e.g. exposure, noise reduction, and sharpening, might even counteract each other. Imaging devices increase the sensitivity of the sensor to have brighter exposure when shooting a low light scene since increasing just exposure time would cause motion blur. However, this process also increases noise since the signal-to-noise ratio diminishes at the same rate as the sensitivity of the sensor increases. To counteract this noise, a second step for the ISP is to apply some sort of local averaging algorithm to smooth areas where noise is most visible. Noise reduction, unfortunately, also causes loss of detail and sharpness. To offset this loss of detail, the third step for an ISP is to process the image with a sharpening algorithm that increases the edge contrast by identifying all edges on the scene and slightly lightening one side of the edges and darkening the other, highlighting the edges and creating an illusion of a sharper image. If any of these steps are not balanced within and between all steps, it will cause multiple overlapping artifacts and degradations, resulting in a loss of perceived quality for the observer. The example above did not take into account white balance and color tuning, nor did it consider any compromises needed to make this processing fast enough for the end user to have a fluent experience of taking photographs. This study included ISPs from the high end to the low end, and, as such, the images represent the general variation in technical quality that is commonly seen in everyday life.

Images were scaled to a size of 1600 × 1200 pixels using the bicubic interpolation method. The four studies consisted of 30 image contents, 27 of which were individual scenes. Image contents were inspired by the photospace, an empirical data of mobile phone camera use frequency mapped against illumination and subject distance from the images (Hultgren & Hertel, 2008) (Figure 1).

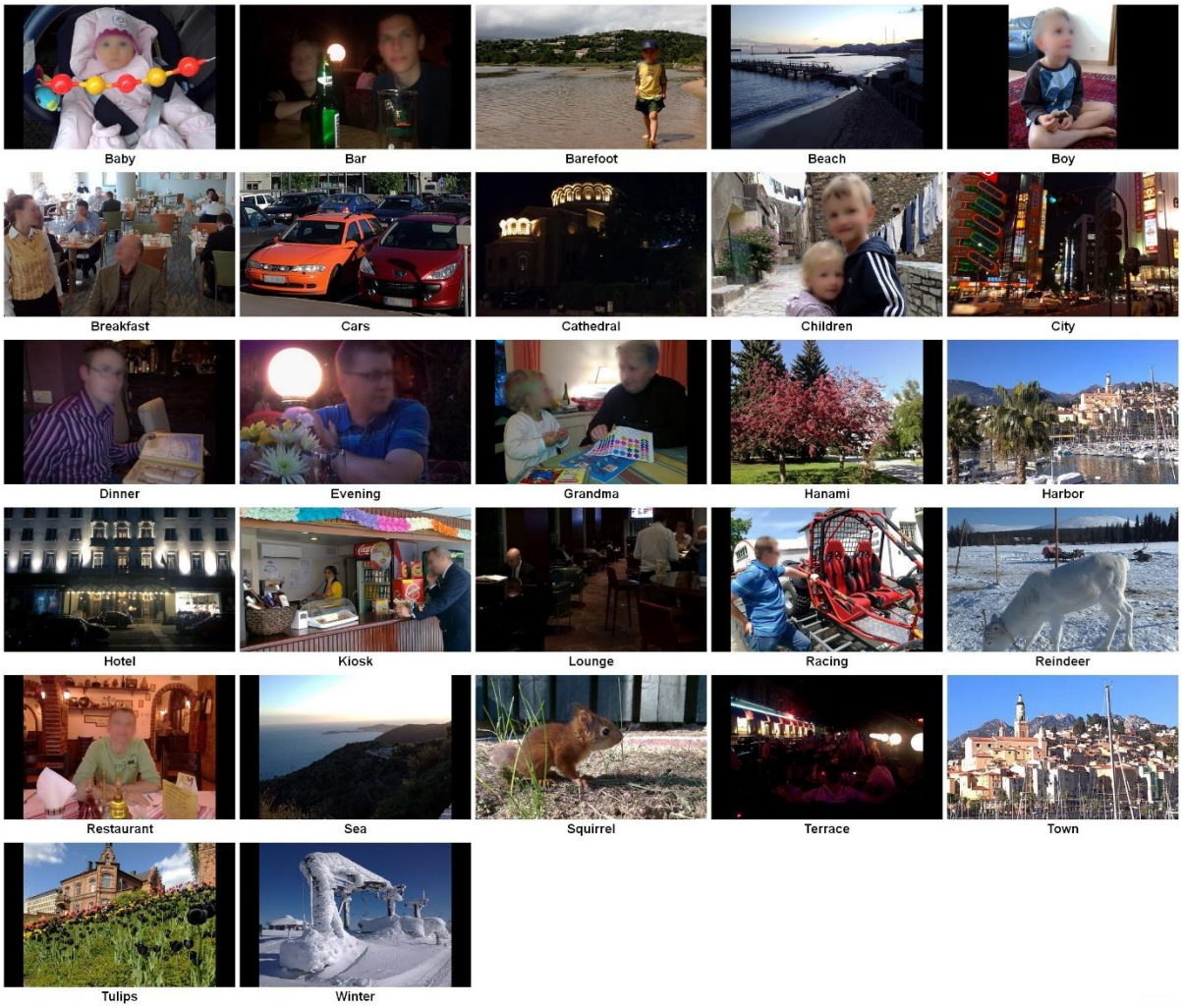

Figure 1. Scenes in the four studies. Faces are blurred for privacy. Original images did not have blurred faces.

Test environment and calibrations

All experiments were conducted in a laboratory setting where the walls had been covered with medium gray curtains to diffuse the ambient illumination. Fluorescent lights (5800 K) were positioned behind the displays and reflected from the gray curtain to create dim and

uniform ambient illumination in the room. The light hitting the displays measured below 20 lx. The observers' viewing distance (~80 cm) was controlled by a line hanging from the ceiling. Observers were instructed to keep their forehead steady next to the line so that their eyes were at the intersection of all three displays. The laboratory had two identical setups so that two observers could conduct the test simultaneously (Figure 2).

Figure 2. Experimental setup and conditions.

Images were presented using three Eizo ColorEdge CG241W displays with 1920 × 1200 pixel resolution, calibrated to sRGB using EyeOne Pro calibrator (X-rite Co.). The target values were 80 cd⁄m2, 6500 K, and gamma 2.2. The consistency of the display calibrations were validated using Photo Research PR-670 spectroradiometer measured at the center of the screen. Figure 3 shows the characterization measurements of displays in test setup 1. Measures from test setup 2 did not differ from test setup 1. The smaller fourth display below the other displays was not characterized as it merely contained a gray graphical user interface

with sliders and text boxes to gather observers' answers. The test protocol also randomized which image was shown on which display to counterbalance potential slight differences in viewing distance, angle, or calibration.

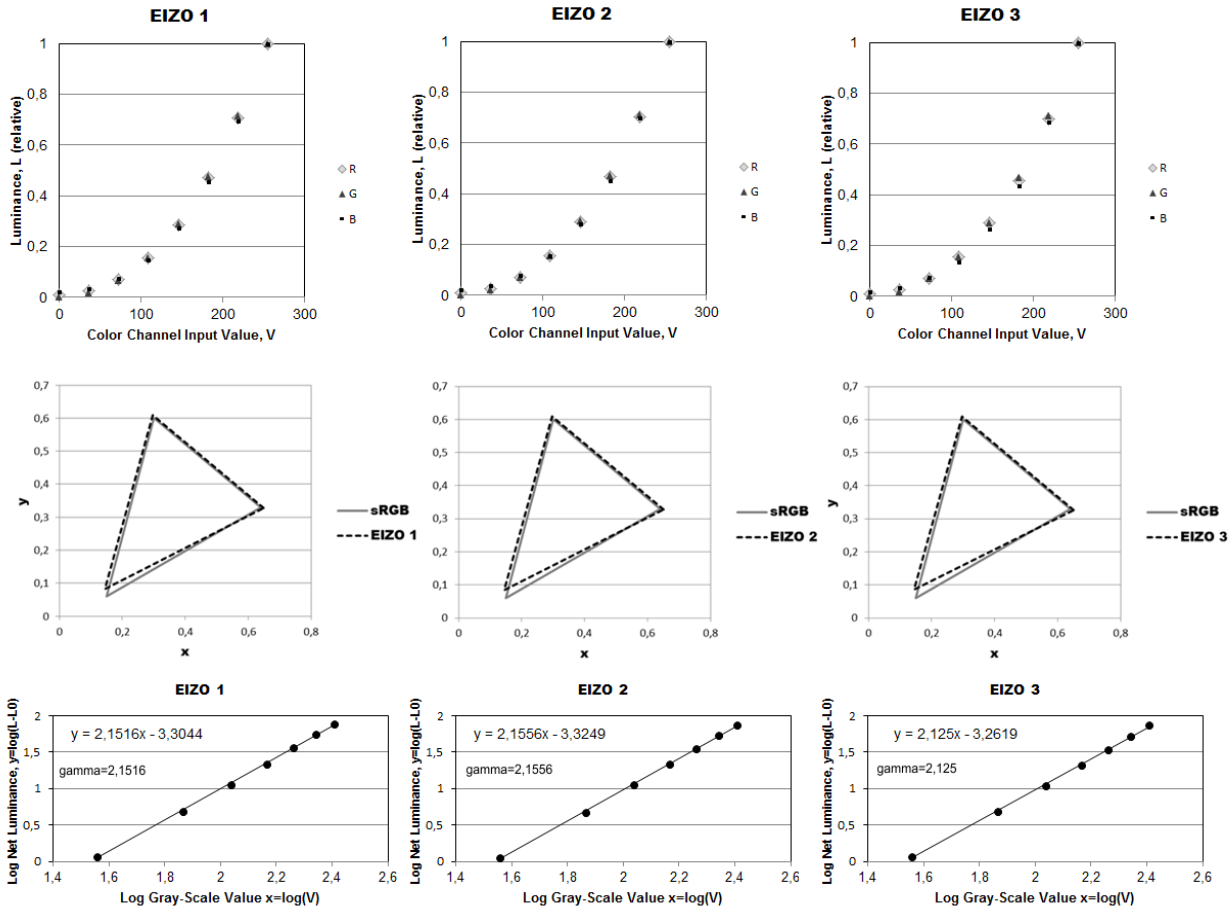

Figure 3. Display characterizations from setup 1. 1st row, luminance measurements of the primary colors. 2nd row, chromaticity coordinates (gamut). 3rd row, electro optical transfer function ("gamma") from gray level luminance values. The gamma value is determined by plotting the log of net luminance L - Lb, where Lb is the luminance of black, against the log of the gray level bit value V. The slope of this log-log plot is "gamma".

Procedures

After providing instructions, the observers conducted practice rounds until they indicated that they were ready for the actual test. They conducted one additional round of evaluation

under supervision of the instructor to confirm that they had understood the instructions. The observers were given the opportunity to have as many breaks as they wanted during the experiment. The average duration of the experiments was 1 h 33 min, the longest being 2 h 20 min and the shortest 55 min. The durations include vision tests, instructions, practice rounds, and breaks. All experiments were conducted using the VQone MATLAB toolbox, a graphical experiment builder for image and video quality evaluations (Nuutinen, Virtanen, Rummukainen, & Häkkinen, 2016).

Triplet comparison

The studies followed a modified soft copy version of the ISO 20462-2 Triplet comparison method (International Organization for Standardization, 2005), where observers saw three images depicting the same scene on separate displays and had a fourth display in which the rating scales were shown. Their task was to rate each image on a scale from 0 to 10. Giving the same score to two images in a triplet was prevented to allow further analysis by ranking the images by their ratings within each triplet. This process did not require any effort from the observer. As an example, if the observer had rated the images as having a score of 6, 9, and 2 from left to right on a single triplet, these could later be ranked as 2nd, 1st, and 3rd for additional analyses.

The randomization protocol in the triplet comparison method used balanced incomplete block (BIB) design, where each stimuli is paired against each other at least once in all of the triplet combinations (Burton & Nerlove, 1976; International Organization for Standardization, 2005). For example, with items 1 to 9, combinations without duplication can be achieved with just 12 triads (1, 2, 4), (4, 5, 7), (7, 8, 1), (2, 3, 5), (5, 6, 8), (8, 9, 2), (1, 3, 6), (4, 6, 9), (7, 9, 3), (1, 5, 9), (4, 8, 3), and (7, 2, 6). Without balancing the blocks and preventing duplicate pairs, nine items would create a complete set of 84 triads, which would create an exhausting experiment for the observers. Note that there were other balanced design

triad combinations besides the one example presented here, which were also randomized for each observer.

Using triplet comparison instead of pair comparison has the benefit of reducing the experiment time since it reduces the number of sample combinations. While the number of sample combinations for paired comparison N is expressed by

$$N = n(n - 1)/2, \quad (1)$$

the number of sample combinations for triplet comparison N is expressed by

$$N = n(n-1)/6\,, \quad (2)$$

where n is the number of samples and n = 2, 3, 4, 5, etc.

As can be seen from Eq. 1 and Eq. 2, triplet comparison with balanced incomplete block (BIB) design reduces the amount of sample combinations to one-third of that of paired comparison. As in our previous nine-sample example, the same sample combinations can be presented with just 12 triplets, while it would require 36 pairs when using paired comparison method. However, not all sample sizes are valid for balanced design triplet comparison without duplicated pairs and the number of samples are restricted to n= 7, 9, 13, 15, 19, 21, 27. Sample sizes of more than 27 are possible, but 27 samples create already 117 triads.

Free descriptions

In addition to the ratings, observers were also were instructed to *"Write down free descriptions for each image and provide the image elements on which you base your judgment. You don't need to use whole sentences."* Using as open instructions as possible, we tried not to influence the observers in any way, as it has been shown that it can have impact on the way people look at an image (Radun et al., 2016; J. Redi et al., 2011). Since the only difference between the images was in technical image quality and the composition and

content remained the same, the free descriptions were reduced to visible image features, such as *sharpness,* and the associations derived from these features, such as *cold* or *gloomy*.

The Triplet comparison method forced the observers to always rate one image as the best and one image as the worst in a set of three images, and then use free descriptions to explain their judgment. We estimated that this method would provide equal distribution of positive and negative valence descriptions about the images, and not skew the valence distribution because of the evaluation task. For example, with a triplet consisting of three very low-quality images, one of them still had to be chosen as the best out of three and the reasoning behind the judgment provided, forcing the observer to find something positive about it.

This study builds on earlier research that comprised a large dataset of 39,415 individual quotations related to aesthetic preference of images from 146 observers (Virtanen et al., 2019). The frequency distribution of the free descriptions from Studies 1-4 has been reported in a previous study exploring whether observers use different terminology when evaluating printed images or digital images. The earlier study also presented an image quality lexicon in the form of an image quality reference wheel.

This publication expands from the earlier study by linking the free descriptions to concurrent image quality ratings and provides an importance rating for each attribute for its effect on the image quality rating. Although having both ratings and the free descriptions in the same experiment made the experiments somewhat longer and more tedious for the observers, it was crucial to gather the free descriptions at the same time as the ratings to determine the aspects of the image on which the evaluation of image quality is based. Earlier studies have also shown that concurrent verbal explanations can enhance stability and reduce variance in image quality ratings (Tuomas Leisti, Radun, Virtanen, Nyman, & Häkkinen, 2014).

Gathering ratings and free descriptions on separate occasions would have severed the link between verbal descriptions and quality ratings as a test strategy. Moreover, observer fatigue and attention could have varied between the tasks. Earlier studies have also shown that the task can have an impact on the way people view and rate an image (Radun et al., 2016; J. Redi et al., 2011).

Observers' free descriptions, e.g. *"very bright, but blurry image"*, were aggregated in a two-step process. First, the grammatical nuances and different inflections, e.g. the terms *bright*, *brighter*, and *brightest*, were all summed up manually under the term *bright*. Second, the remaining terms were cross-referenced for synonyms, e.g. *bright*, *luminous*, and *radiant*, to form the final attribute *bright*. Synonyms were identified using FinnWordNet version 2.0 lexical database for Finnish, a derivative of the Princeton WordNet 3.0, where the original English words had been translated to Finnish by professional translators (Linden & Carlson, 2010). FinnWordNet contains words (nouns, verbs, adjectives, and adverbs) grouped by meaning into synonym groups representing concepts. These synonym groups are linked to each other with such relations as hyponymy and antonymy, creating a semantic network. For example, the final attribute of *loud colors* contains the following quotations: *harsh colors, screaming colors, loud colors, very loud colors, sparkling colors, quite loud colors, slightly too loud colors, absolutely too loud colors, a little bit too loud colors, unnaturally loud colors, nice loud colors, colors hurts your eyes, colors jump out too much.* As the original quotations were in Finnish, *loud* in this context could also be translated as *glaring, blatant, garish, gaudy, flagrant, egregious, flashy*, or *brash*. A more detailed description of the procedure is provided in Virtanen et al. (2019).

Results

Data screening

There were no missing data in the preference evaluations, nor were there inaccurate values outside the requested 0 to 10 MOS scale, as observers' preferences were recorded using a VQOne Matlab program that prevented omission of answers or using inappropriate values outside the given scale (Nuutinen et al., 2016).

Figure 4 shows that the preference Mean Opinion Score (MOS) follows normal distribution quite well. Multivariate outlier values were identified using Mahalanobis distance with linear regression analysis. The Mahalanobis distance was compared against a 99.9% threshold at df=68 of the Chi-squared distribution table. Using this method, 15.6% of the values were flagged as multivariate outliers, e.g. having a combination of attributes and MOS that deviates from the overall average of averages. However, upon further inspection 91% of the flagged multivariate outliers were from attributes with frequency of less than 200 quotations. Some of the multivariate outliers contained all quotations from a single attribute and removing them would mean that potentially interesting data might be omitted from the analysis. As there are no right or wrong answers in subjective image quality preference judgments, we felt that removing them as outliers was not warranted. Therefore, we decided to leave them in the analysis.

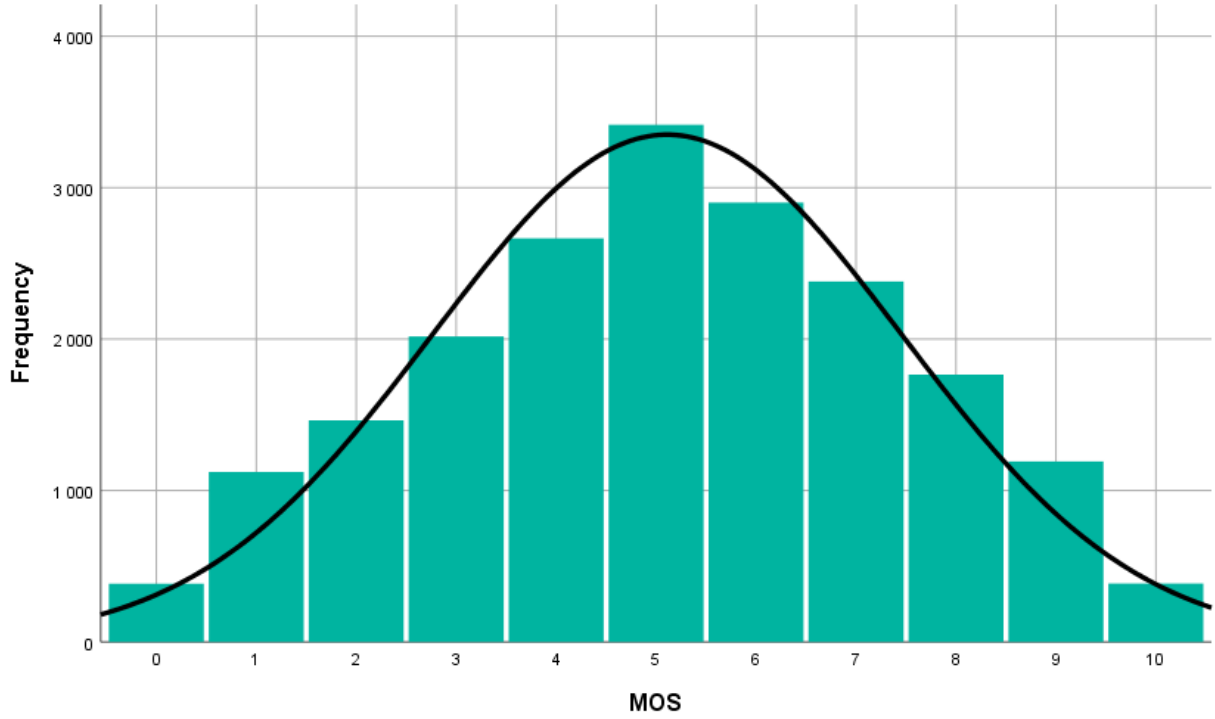

Figure 4. Mean Opinion Score (MOS) distribution across all four studies.

Of the 19,692 preference judgments, 5.4% did not have any verbal description attached to them. No systematic pattern was found for the missing verbal descriptions in the data. Of the remaining evaluations, 58.0 % had verbal descriptions that could be aggregated into one attribute, 30.1% yielded two attributes per verbal description, and 5.4% yielded three attributes. The remaining 1.0% had more than three attributes, six attributes being the limit.

As one-third of the evaluations had more than one attribute, we could examine which attributes had the most co-occurrences together. As we were only interested in cases where attributes were present instead of absent, we used Jaccard's similarity coefficient for binary variables that excludes joint absences from the analysis (Jaccard, 1912). This means that only joint occurrences were considered as an indication of similarity, whereas joint absences were omitted. Jaccard's similarity index can have values from 0 to 1 that can be interpreted as a percentage of co-occurrences. The strongest similarity was observed with *Unnatural* and *Unreal* attributes, where Jaccard's similarity was 0.504, meaning that these two attributes had

half of their occurrences together in cases where they were not occurring alone. However, these joint occurrences between *Unnatural* and *Unreal* only amounted to 0.9% of the total combined frequency of the two attributes, making the multicollinearity insignificant. Other attributes had weaker relationships, and the second highest Jaccard's similarity coefficient was only 0.098 between the attributes *Sharp* and *Too dark colors* (Table 2).

Table 2

Ten attributes with the most co-occurrences.

Attribute 1	**Attribute 2**	**Jaccard's similarity**
Unnatural	Unreal	0.504
Sharp	Too dark colors	0.098
Pastel colors	Photoshopped	0.095
Natural	Unsharp	0.094
Too bright	Yellow	0.092
Purple	Noisy	0.092
Photoshopped	Real colors	0.091
Faded colors	Overexposed	0.090
Photoshopped	Soft	0.090
Sharp	Too bright	0.090

Figure 5 compares the ranking data of the triplets to the attributes. It shows the frequency distribution of each attribute whether it was mentioned when ranking the image as best out of three (Rank 1), in between (Rank 2), or worst out of three (Rank 3). Attributes are sorted so that attributes linked most often to the best out of three (Rank 1) are at the top, while attributes linked most often to the worst out of three (Rank 3) are at the bottom.

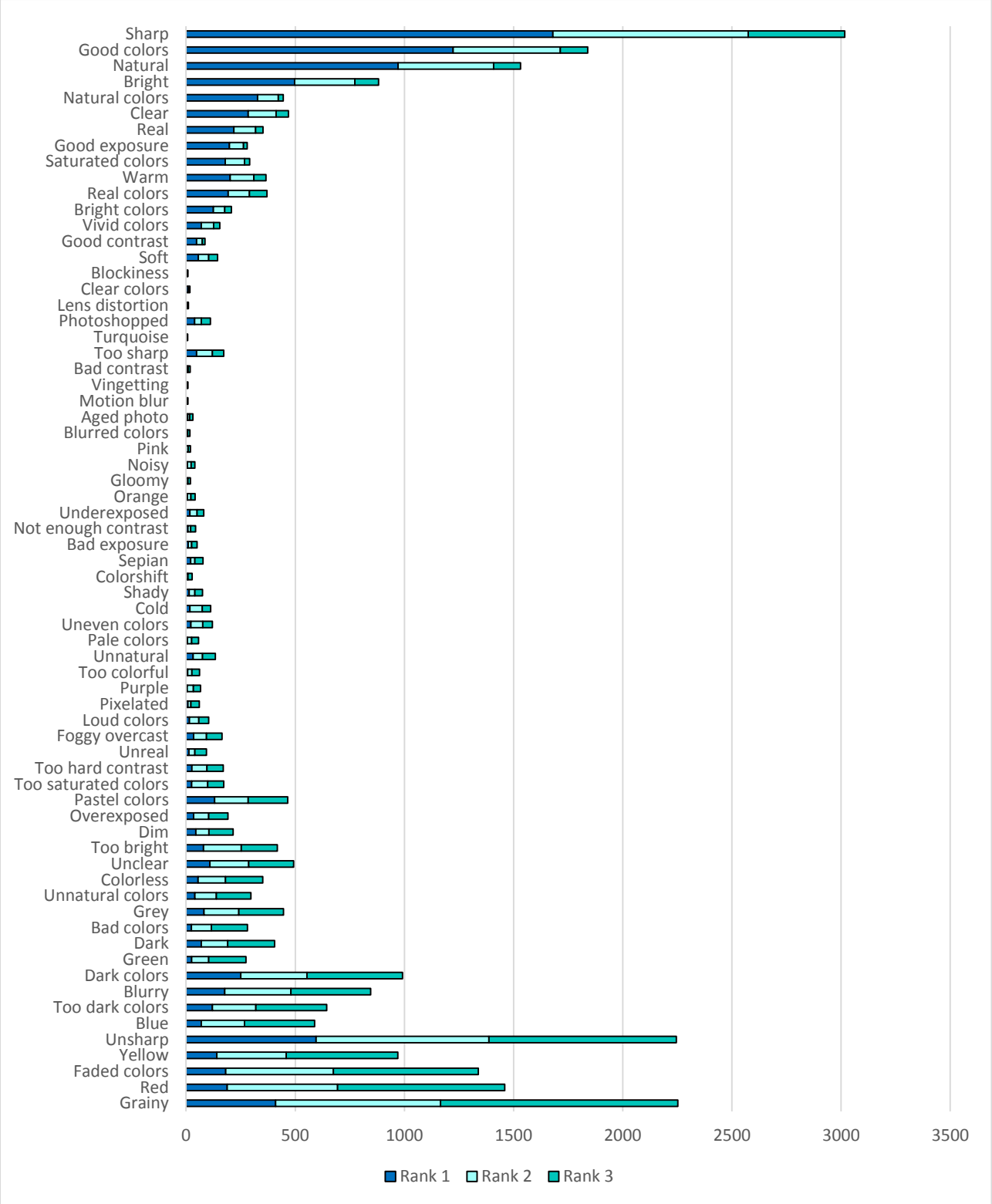

Figure 5. Stacked attribute frequency separated by ranking of the images in each triplet. The attribute order is sorted so that attributes linked most often to the best out of three (Rank 1) are at the top, while attributes linked most often to the worst out of three (Rank 3) are at the bottom.

Linear regression modeling

We entered all 68 attributes in the linear regression model as predictors and used MOS as a dependent variable $R^2 = 0.427$, $F(68,19623) = 214.691$, $p < 0.000$. The 10 attributes *Bad contrast, Bad Exposure, Blockiness, Blurred colors, Lens distortion, Not enough contrast, Too colorful, Too saturated colors, Too sharp, and Turquoise* were not statistically significant predictors (Table 3). To get an importance value for each predictor attribute, we used the leave-one-out method, based on the residual sum of squares (SSe) by removing one predictor at a time from the final full model and normalizing the extracted predictor importance to have a cumulative percentage value of 100% (International Business Machines Corporation, 2017). Table 3 presents the importance of each attribute. The higher the importance, the more influence that attribute has on the predictive model. For example, the attributes *Grainy*, *Sharp*, and *Natural* influence together 36% of the predictive power of the final regression model. Mean Opinion Score (MOS) value for each attribute represents the average image quality of all of the images that have been commented to have that attribute. For example, if the image has been noted to be *Bright* it would get an average image quality score of 7.41 or to have *Unnatural colors* it would get a score of 4.34. We can also evaluate the valence of each attribute by examining the regression coefficient B of the linear regression model. Negative values decrease the predicted image quality and have a negative valence, while positive values increase it and have a positive valence. The further the value is from zero, the stronger the effect on preference estimation.

Table 3

Attribute frequency, regression coefficients, importance, and mean opinion scores (MOS).

The table is sorted by importance, showing the most influential attributes first.

		Unstandardized Coefficients		Standardized Coefficients					
Predictor	N	B	Std. Error	Beta	t	Sig.	Importance (%)	MOS	Adj Pred.
(Constant)		5.337	0.029		184.871	0.000			
Grainy	2252	-1.723	0.042	-0.234	-41.222	0.000	14.55	3.67	3.45
Sharp	3017	1.379	0.037	0.212	36.956	0.000	11.70	7.11	6.90
Natural	1532	1.654	0.050	0.189	33.276	0.000	9.48	7.26	7.04
Red	1459	-1.401	0.051	-0.157	-27.416	0.000	6.44	3.88	3.64
Bright	882	1.678	0.063	0.148	26.776	0.000	6.14	7.41	7.23
Good Colors	1838	1.130	0.046	0.14	24.820	0.000	5.28	6.79	6.61
Unsharp	2246	-1.029	0.042	-0.139	-24.613	0.000	5.19	4.41	4.21
Dark Colors	991	-1.318	0.060	-0.123	-22.121	0.000	4.19	3.92	3.74
Blurry	845	-1.330	0.064	-0.115	-20.688	0.000	3.67	3.85	3.66
Real Colors	370	1.880	0.095	0.109	19.890	0.000	3.39	6.56	6.45
Unclear	492	-1.515	0.083	-0.101	-18.319	0.000	2.87	3.72	3.57
Clear	469	1.434	0.084	0.093	17.005	0.000	2.48	6.89	6.79
Yellow	970	-0.984	0.061	-0.091	-16.190	0.000	2.24	4.44	4.26
Bad Colors	281	-1.741	0.108	-0.088	-16.054	0.000	2.21	3.62	3.52
Saturated Colors	291	1.485	0.106	0.076	14.027	0.000	1.68	7.26	6.91
Blue	589	-0.984	0.076	-0.072	-12.902	0.000	1.43	4.21	3.99
Gray	446	-1.075	0.087	-0.068	-12.419	0.000	1.32	4.48	4.18
Dark	405	-1.092	0.091	-0.066	-11.967	0.000	1.23	4.44	4.20
Natural Colors	445	0.979	0.086	0.062	11.318	0.000	1.10	6.49	6.16
Green	274	-1.233	0.109	-0.062	-11.267	0.000	1.09	3.76	3.59
Vivid Colors	154	1.632	0.145	0.061	11.287	0.000	1.09	6.83	6.63
Warm	366	1.070	0.095	0.062	11.300	0.000	1.09	6.52	6.17
Too Dark Colors	644	-0.774	0.073	-0.059	-10.565	0.000	0.96	4.94	4.74
Faded Colors	1338	-0.542	0.053	-0.058	-10.271	0.000	0.90	5.06	4.84
Loud Colors	104	-1.749	0.176	-0.054	-9.959	0.000	0.85	3.52	3.44
Real	352	0.962	0.097	0.054	9.885	0.000	0.84	6.70	6.56
Bright Colors	208	1.120	0.125	0.049	8.957	0.000	0.69	6.62	6.43
Unnatural Colors	297	-0.952	0.110	-0.049	-8.626	0.000	0.64	4.34	4.07
Pixelated	61	-1.562	0.229	-0.037	-6.832	0.000	0.40	2.92	2.66
Colorless	351	-0.661	0.098	-0.037	-6.779	0.000	0.39	4.94	4.56

Uneven Colors	121	-1.093	0.163	-0.036	-6.693	0.000	0.38	4.07	4.00
Dim	216	-0.773	0.123	-0.034	-6.295	0.000	0.34	4.55	4.50
Overexposed	192	-0.812	0.130	-0.034	-6.264	0.000	0.34	4.68	4.55
Too Hard Contrast	170	-0.861	0.138	-0.034	-6.243	0.000	0.33	4.80	4.54
Unnatural	134	-0.906	0.157	-0.032	-5.764	0.000	0.28	4.37	4.07
Purple	66	-1.178	0.221	-0.029	-5.329	0.000	0.24	3.08	2.80
Clear Colors	17	2.189	0.432	0.027	5.062	0.000	0.22	7.18	7.00
Good Exposure	280	0.530	0.108	0.027	4.894	0.000	0.21	6.67	6.31
Unreal	93	-0.936	0.195	-0.027	-4.809	0.000	0.20	3.69	3.55
Foggy Overcast	164	-0.655	0.140	-0.025	-4.667	0.000	0.19	5.07	4.76
Orange	41	-1.264	0.279	-0.025	-4.536	0.000	0.18	4.05	3.75
Sepian	77	-0.896	0.203	-0.024	-4.424	0.000	0.17	4.51	4.35
Noisy	40	-1.171	0.282	-0.022	-4.148	0.000	0.15	3.88	3.77
Cold	113	-0.685	0.169	-0.022	-4.062	0.000	0.14	5.08	4.71
Pale Colors	57	-0.972	0.237	-0.022	-4.108	0.000	0.14	4.40	4.19
Too Bright	418	-0.336	0.089	-0.021	-3.756	0.000	0.12	5.40	5.14
Soft	144	0.509	0.149	0.018	3.406	0.001	0.10	5.97	5.69
Vingetting	8	-2.152	0.630	-0.018	-3.417	0.001	0.10	2.50	2.50
Good Contrast	86	0.571	0.193	0.016	2.960	0.003	0.08	6.48	6.18
Pastel Colors	465	-0.251	0.084	-0.016	-2.969	0.003	0.08	5.31	5.15
Colorshift	28	-0.958	0.337	-0.015	-2.841	0.004	0.07	4.39	4.07
Gloomy	20	-1.081	0.399	-0.015	-2.711	0.007	0.06	4.20	3.90
Aged Photo	31	0.790	0.321	0.013	2.462	0.014	0.05	5.16	5.10
Motion Blur	9	-1.408	0.594	-0.013	-2.373	0.018	0.05	3.67	3.67
Shady	75	0.488	0.206	0.013	2.366	0.018	0.05	6.15	5.84
Photoshopped	111	-0.349	0.172	-0.011	-2.033	0.042	0.04	5.10	4.81
Pink	20	-0.820	0.399	-0.011	-2.058	0.040	0.04	4.30	4.00
Underexposed	81	-0.422	0.201	-0.012	-2.099	0.036	0.04	5.23	4.96
Bad Exposure	50	-0.496	0.258	-0.011	-1.926	0.054	0.03	4.64	4.28
Too Sharp	172	-0.267	0.137	-0.011	-1.947	0.052	0.03	5.43	5.14
Bad Contrast	19	-0.664	0.409	-0.009	-1.623	0.105	0.02	4.58	4.26
Blurred Colors	18	-0.674	0.420	-0.009	-1.605	0.109	0.02	5.39	5.28
Lens Distortion	11	-0.663	0.538	-0.007	-1.233	0.218	0.01	4.91	4.90
Blockiness	9	0.107	0.596	0.001	0.180	0.857	0.00	4.11	4.11
Not Enough Contrast	44	-0.060	0.269	-0.001	-0.225	0.822	0.00	5.30	5.14
Too Colorful	62	-0.143	0.227	-0.003	-0.631	0.528	0.00	5.40	5.14
Too Saturated Colors	172	-0.057	0.138	-0.002	-0.413	0.680	0.00	5.65	5.17
Turquoise	7	-0.427	0.676	-0.003	-0.631	0.528	0.00	3.71	3.71

Figure 6 indicates that our linear regression model seems to have slight heteroscedasticity. With low Mean Opinion Score (MOS) values, prediction error deviates upwards, while with high MOS values prediction error deviates downwards. This means that with low quality the prediction overestimates the values and with high quality the prediction underestimates the quality compared with the predicted quality evaluation.

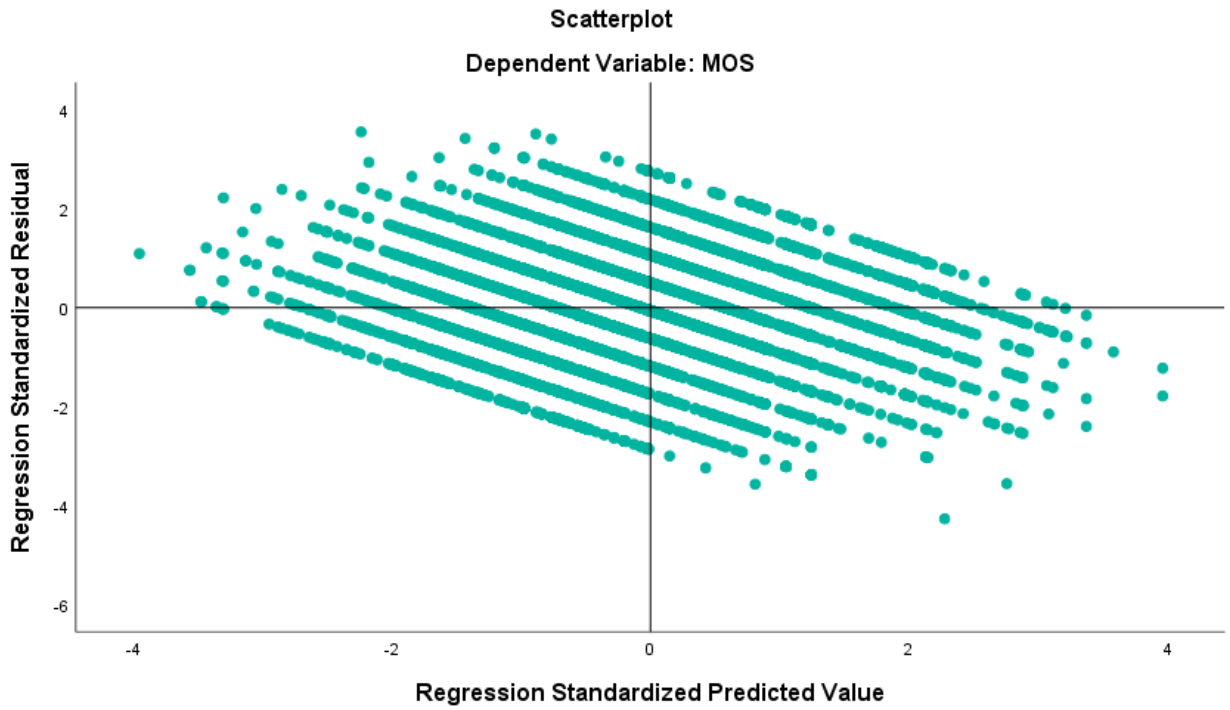

Figure 6. Standardized residual and predicted value.

Discussion

This study examined the words observers use when evaluating image quality. Determining how various elements, such as sharpness or naturalness, affect image quality can shed light on the processes behind the image quality evaluation task. These results support the suggestion made by Nyman et al. (2010) that the subjective decision space can change as a function of preference. Figure 5 shows that observers use different terms when describing high-quality images and low-quality images. Results also suggest that there could be some sort of high-level and low-level distinction to be made in observer description of image

elements, as Leisti et al. (2009) proposed earlier. The high-level and low-level distinction can be made mostly from the higher quality images, while low-quality images are mostly only described by more concrete image fidelity-related elements such as graininess, color cast, lack of sharpness, and exposure issues (Figure 5). Certain attributes, such as *Brightness*, *Naturalness*, or *Good colors*, seem to be related to with high image quality. However, when looking at Table 3, the most important attributes are G*rainy, Sharp, Natural, Bright, Red,* and *Unsharp.* Besides the single attribute *Natural,* the other attributes seem to be related to image fidelity. This could indicate that a certain level of image fidelity must be achieved before more subjective higher level elements, such as naturalness and others, can emerge. At least in the case of photographs, processing fluency therefore has an effect on the perception of aesthetic pleasure, as noted elsewhere (Reber et al., 2004). Painters, in general, have a much greater freedom in creating their art and conveying their emotions and thoughts with brush strokes on a blank canvas. Photographers, on the other hand, are somewhat more constrained, as their canvas has to exist in the natural world, in one form or another, to be photographed in the first place (Joshi et al., 2014). This result is also quite nicely explained by the dual-process perspective on processing fluency. In particular, abstract art has a very low level of processing fluency, but introduces hedonic pleasure through cognitive enrichment. Natural scene images, in turn, do not trigger cognitive enrichment and are therefore processed on an automatic level, with high processing fluency (Graf & Landwehr, 2015). Observers also seem to understand that the images are meant to be representations of the real world, and therefore, when they appear natural it will have a significant impact on the perceived quality of the image (Tinio et al., 2011).

Even with the effort to balance out the evaluation task effect on negative or positive bias in word use, 72% of the attributes had a negative impact on the preference judgment. Previous studies have shown contradictory results on whether the bias is negative (S.

Yendrikhovskij, MacDonald, Bech, & Jensen, 1999) or positive (Jacobsen, Buchta, Köhler, & Schröger, 2004). This study differed from the earlier studies in that we did not interpret the valence of the attribute merely from the words, but on the effects they had as predictors on the preference judgments (Table 3). This way of revealing the valence of the attributes could bring deeper insight since some attributes had unexpected valence that might not have otherwise been found. For example, the attributes *Soft* and *Shady* had a positive valence. A word of caution is still needed when using the regression coefficients as an indication of valence; critical evaluation case by case is recommended. For example, the attribute *Blockiness* has a positive regression coefficient, but also has very small frequency, which significantly decreases its reliability for the information offered. In this study, the overall negative bias can arise from there being more ways that observers can perceive images to fail than to excel when evaluating image quality. Another explanation could be that observers lack exact words and do not comment when some image-degrading element, such as graininess, is missing.

A limitation of this study is that almost all of the observers were female students. An even more severe issue for generalization of the results comes from the fact that all participants had native fluency in Finnish language or were native speakers. The Finnish language is part of the Finno-Ugric family and differs from, for instance, the English language, which is an Indo-European family. Differences in language and cultural background might have some effect on how observers describe the images that they rate. Future work should replicate the results with different cameras or ISPs and observers from different cultural backgrounds and with a more even gender distribution.

References

Augustin, M. D., Wagemans, J., & Carbon, C.-C. C. (2012). All is beautiful? Generality vs. specificity of word usage in visual aesthetics. *Acta Psychologica, 139*(1), 187–201.

https://doi.org/10.1016/j.actpsy.2011.10.004

Bech, S., Hamberg, R., Nijenhuis, M., Teunissen, K., Looren de Jong, H., Houben, P., … SPIE. (1996). Rapid perceptual image description (RaPID) method. In *Proc. SPIE 2657, Human Vision and Electronic Imaging* (Vol. 2657, pp. 317–328). San Jose, CA, United States: Society of Photo-Optical Instrumentation Engineers (SPIE). https://doi.org/10.1117/12.238728

Berlyne, D. E. (1972). Uniformity in variety: extension to three-element visual patterns and to non-verbal measures. *Canadian Journal of Psychology*, *26*(3), 277–291. https://doi.org/10.1037/h0082436

Burton, M. L., & Nerlove, S. B. (1976). Balanced designs for triads tests: Two examples from English. *Social Science Research*, *5*(3), 247–267. https://doi.org/10.1016/0049-089X(76)90002-8

Callet, P. Le, Möller, S., & Perkins, A. (2012). *Qualinet White Paper on Definitions of Quality of Experience* (Vol. 1.2). Lausanne, Switzerland. Retrieved from www.qualinet.eu

Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., & Bromberger, B. (2010). The Assessment of Art Attributes. *Empirical Studies of the Arts*, *28*(2), 207–222. https://doi.org/10.2190/EM.28.2.f

Chen, B., Rhodes, C., Crawford, A., & Hambuchen, L. (2015). Wineinformatics: Applying data mining on wine sensory reviews processed by the computational wine wheel. *IEEE International Conference on Data Mining Workshops, ICDMW*, *2015-Janua*(January), 142–149. https://doi.org/10.1109/ICDMW.2014.149

Datta, R., & Wang, J. Z. (2010). ACQUINE. In *Proceedings of the international conference*

on Multimedia information retrieval - MIR '10 (p. 421). New York, New York, USA: ACM Press. https://doi.org/10.1145/1743384.1743457

Dodge, S., & Karam, L. (2019). Human and DNN Classification Performance on Images With Quality Distortions : A Comparative Study, *16*(2).

Engeldrum, P. G. (2004). A Theory of Image Quality: The Image Quality Circle. *Journal of Imaging Science and Technology*, *48*(5), 447–457.

Fechner, G. T. (1876). *Vorshule der Aesthetik*. Leipzig: Breitkopf & Härtel.

Fliegel, K. (eds. . K. (2013). QUALINET Multimedia Databases v4.6. Retrieved May 12, 2014, from www.qualinet.eu

Gawel, R., Oberholster, A., & Francis, I. L. (2000). A 'Mouth-feel Wheel': terminology for communicating the mouth-feel characteristics of red wine. *Australian Journal of Grape and Wine Research*, *6*(3), 203–207. https://doi.org/10.1111/j.1755-0238.2000.tb00180.x

Graf, L. K. M., & Landwehr, J. R. (2015). A Dual-Process Perspective on Fluency-Based Aesthetics: The Pleasure-Interest Model of Aesthetic Liking. *Personality and Social Psychology Review*, *19*(4), 395–410. https://doi.org/10.1177/1088868315574978

Hultgren, B. O., & Hertel, D. W. (2008). Megapixel mythology and photospace: estimating photospace for camera phones from large image sets. *Proc. SPIE*, *6808*(January 2008), 680810–680818. https://doi.org/10.1117/12.766611

I3A. (2007). *CPIQ Initiative Phase 1 White Paper: Fundamentals and review of considered test methods*. Retrieved from http://www.i3a.org/resources/cpiq/

International Business Machines Corporation. (2017). IBM SPSS Modeler 18.1.1 Algorithms Guide. Retrieved October 3, 2019, from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.1.1/en/Algo

rithmsGuide.pdf

International Organization for Standardization. (2005). *ISO 20462-2 - Photography - Psychophysical experimental methods for estimating image quality - Part 2: Triplet comparison method* (Vol. 2).

Jaccard, P. (1912). the Distribution of the Flora in the Alpine Zone. *New Phytologist*, *11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jacobsen, T., Buchta, K., Köhler, M., & Schröger, E. (2004). The Primacy of Beauty in Judging the Aesthetics of Objects. *Psychological Reports*, *94*(3_suppl), 1253–1260. https://doi.org/10.2466/pr0.94.3c.1253-1260

Janssen, T. J. W. M., & Blommaert, F. J. J. (1997). Image Quality Semantics. *Journal of Imaging Science and Technology*, *41*(5), 555–560. https://doi.org/10.1117/3.pm101.ch3

Joshi, D., Datta, R., Fedorovskaya, E., Lu, X., Luong, Q.-T., Wang, J. Z., … Luo, J. (2014). On Aesthetics and Emotions in Scene Images: In *Scene Vision* (pp. 241–272). The MIT Press. https://doi.org/10.7551/mitpress/9780262027854.003.0012

Keelan, B. (2002). *Handbook of image quality: characterization and prediction*. CRC Press.

Lawless, L. J. R., & Civille, G. V. (2013). Developing Lexicons: A Review. *Journal of Sensory Studies*, *28*(4), 270–281. https://doi.org/10.1111/joss.12050

Lawless, L. J. R., Hottenstein, A., & Ellingsworth, J. (2012). The McCormick spice wheel: A systematic and visual approach to sensory lexicon development. *Journal of Sensory Studies*, *27*(1), 37–47. https://doi.org/10.1111/j.1745-459X.2011.00365.x

Leder, H. (2014). 6 Beyond perception – information processing approaches to art appreciation. In *The Cambridge Handbook of the Psychology of Aesthetics and the Arts* (pp. 115–138).

Leder, L., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, *95*(4), 489–508. https://doi.org/http://doi.org/10.1348/0007126042369811

Leisti, T., Radun, J., Virtanen, T., Halonen, R., & Nyman, G. (2009). Subjective Experience of image quality: Attributes, definitions and decision making of subjective image quality. In S. P. Farnand & F. Gaykema (Eds.), *Proceedings of SPIE - The International Society for Optical Engineering* (Vol. 7242). San Jose, CA, United States. https://doi.org/10.1117/12.807142

Leisti, Tuomas, Radun, J., Virtanen, T., Nyman, G., & Häkkinen, J. (2014). Concurrent explanations can enhance visual decision making. *Acta Psychologica*, *145*(1), 65–74. https://doi.org/10.1016/j.actpsy.2013.11.001

Linden, K., & Carlson, L. (2010). FinnWordNet – WordNet på finska via översättning. *LexicoNordica*, *17*, 119–140. Retrieved from http://www.ling.helsinki.fi/~klinden/pubs/FinnWordnetInLexicoNordica-en.pdf

Mayer, S., & Landwehr, J. R. (2018). Quantifying visual aesthetics based on processing fluency theory: Four algorithmic measures for antecedents of aesthetic preferences. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(4), 399–431. https://doi.org/10.1037/aca0000187

Meilgaard, M. C., Dalgliesh, C. E., & Clapperton, J. F. (1979). Beer Flavour Terminology. *Journal of the Institute of Brewing*, *85*(1), 38–42. https://doi.org/10.1002/j.2050-0416.1979.tb06826.x

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-Reference Image Quality Assessment in the Spatial Domain. *Image Processing, IEEE Transactions On*, *21*(12), 4695–4708. https://doi.org/10.1109/TIP.2012.2214050

Nuutinen, M., Virtanen, T., Rummukainen, O., & Häkkinen, J. (2016). VQone MATLAB toolbox: A graphical experiment builder for image and video quality evaluations. *Behavior Research Methods*, *48*(1), 138–150. https://doi.org/10.3758/s13428-014-0555-y

Nyman, G., Häkkinen, J., Koivisto, E.-M., Leisti, T., Lindroos, P., Orenius, O., … Vuori, T. (2010). Evaluation of the visual performance of image processing pipes: Information value of subjective image attributes. In *Proceedings of SPIE - The International Society for Optical Engineering* (Vol. 7529). https://doi.org/10.1117/12.839946

O'Hare, D. P. A., & Gordon, I. E. (1977). Dimensions of the perception of art: verbal scales and similarity judgements. *Scandinavian Journal of Psychology*, *18*(1), 66–70. https://doi.org/10.1111/j.1467-9450.1977.tb00257.x

Pedersen, M., Bonnier, N., Hardenberg, J. Y., & Albregtsen, F. (2010). Attributes of image quality for color prints. *Journal of Electronic Imaging*, *19*(1), 011016. https://doi.org/10.1117/1.3277145

Radun, J., Nuutinen, M., Leisti, T., & Häkkinen, J. (2016). Individual differences in image-quality estimations: Estimation rules and viewing strategies. *ACM Transactions on Applied Perception (TAP)*, *13*(3), 14. https://doi.org/10.1145/2890504

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: *Personality & Social Psychology Review*, *8*(4), 364–382. https://doi.org/https://doi.org/10.1207/s15327957pspr0804_3

Redi, J. A., Zhu, Y., Ridder, de, H., De Ridder, H., & Heynderickx, I. E. J. (2015). How passive image viewers became active multimedia users : new trends and recent advances in subjective assessment of quality of experience. In C. Deng, L. Ma, W. Lin, & K. N. Ngan (Eds.), *Visual signal quality assessment : quality of experience (QoE)* (pp. 31–72).

Germany: Springer. https://doi.org/10.1007/978-3-319-10368-6_2

Redi, J., Liu, H., Zunino, R., & Heynderickx, I. (2011). Interactions of visual attention and quality perception. In *Human Vision and Electronic Imaging XVI. Proc. of SPIE-IS&T Electronic Imaging* (Vol. 7865). San Francisco, (CA). https://doi.org/10.1117/12.876712

Saad, M. A., Bovik, A. C., & Charrier, C. (2010). A DCT Statistics-Based Blind Image Quality Index. *Signal Processing Letters, IEEE*, *17*(6), 583–586. https://doi.org/10.1109/LSP.2010.2045550

Tinio, P. P. L., & Leder, H. (2009). Natural Scenes Are Indeed Preferred, but Image Quality Might Have the Last Word. *Psychology of Aesthetics, Creativity, and the Arts*, *3*(1), 52–56. https://doi.org/10.1037/a0014835

Tinio, P. P. L., Leder, H., & Strasser, M. (2011). Image quality and the aesthetic judgment of photographs: Contrast, sharpness, and grain teased apart and put together. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(2), 165–176. https://doi.org/10.1037/a0019542

To, M., Lovell, P. G., Troscianko, T., & Tolhurst, D. J. (2008). Summation of perceptual cues in natural visual scenes. *Proceedings. Biological Sciences / The Royal Society*, *275*(1649), 2299–2308. https://doi.org/10.1098/rspb.2008.0692

Virtanen, T., Nuutinen, M., & Häkkinen, J. (2019). Image quality wheel. *Journal of Electronic Imaging*, *28*(1). https://doi.org/10.1117/1.JEI.28.1.013015

Yao, L., Suryanarayan, P., Qiao, M., Wang, J. Z., & Li, J. (2012). OSCAR: On-site composition and aesthetics feedback through exemplars for photographers. *International Journal of Computer Vision*, *96*(3), 353–383. https://doi.org/10.1007/s11263-011-0478-3

Yendrikhovskij, S., MacDonald, L., Bech, S., & Jensen, K. (1999). Enhancing colour image

quality in television displays. *The Imaging Science Journal*, *47*(4), 197–211. https://doi.org/10.1080/13682199.1999.11736360

Yendrikhovskij, S. N., De Ridder, H., Fedorovskaya, E. A., & Blommaert, F. J. J. (1997). Colourfulness judgments of natural scenes. *Acta Psychologica*, *97*(1), 79–94. https://doi.org/10.1016/S0001-6918(97)00025-5

Zarzo, M., & Stanton, D. T. (2009). Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps. *Attention, Perception & Psychophysics*, *71*(2), 225–247. https://doi.org/10.3758/APP.71.2.225